NOO CONTRACT TRANS 12

Reprint from the Periodical
NATURAL SCIENCES

Springer Publishing House/ Berlin - Heidelberg - New York
1965 Issue No. 1, pp. 1-26 52th year

AN AUTOMATIC CLASSIFICATION SYSTEM FOR PUBLICATIONS, LIBRARIES AND DOCUMENTATION

by

Martin Scheele, Schlitz/Hessen

Pages translated: 1-26

Contractor: Translation Consultants, Ltd.,
944 South Wakefield St., Room 302
Arlington, Va. 22204

Delivery Date: 30 September 1968

Prepared under Contract #N62306-69-M-0201 for the U.S. Naval Oceanographic Office

AN AUTOMATIC CLASSIFICATION SYSTEM FOR PUBLICATIONS, LIBRARIES AND DOCUMENTATION*

by

Martin Scheele, Schlitz/Hessen

I. INTRODUCTION

p. 1

1. Statement of the Problem

It is the task of documentation to handle a more rapidly produced and ever increasing volume of written materials. In this respect, scientific periodicals are in the foreground of our interest. The magnitude of the task and the volume of literary material force us to rationalize by using modern methods. The documentation problem itself is multifaceted and, therefore, cannot be solved in one operation or on one level. We must, rather, differentiate between various stages of documentation, beginning with the simple processing of the titles of publications, extending through the citing of key words and references all the way to the evaluation of concrete data and values. Therefore, before using any modern methods, we must ask at what stage of documentation the best rationalization effect can be obtained, and which specific method promises the most favorable relationship between expenditures and profit.

A thorough analysis of this question led to the conclusion that the automation of documentation is best begun at the first, bibliographic level of pure title processing, and according to p. 2 experience, it is most rational to use punched tape techniques in combination with electronic computers.

---

*Dedicated with gratitude to the memory of my old friend, Grad. Eng. Willi Heimerdinger, who died on 6 March 1964.

The description of such a system is the subject of this paper. My main objective is to familiarize the reader with the basic thought processes and work sequences of our system. Details are only discussed if they are indispensable for the understanding of the whole. Otherwise I have made every effort to leave out all specifics and complications, particularly since it is necessary to save space.

The described method was developed in the Documentation Center for Biology in Schlitz/Hessen. The programming for the electronic computers was planned and carried out by Mr. Gerhardt Natalis. I would like to take this opportunity to thank him for his work and point out that his accomplishment can only be truly appreciated to its full extent by a programming expert. IBM Germany generously supplied us with the electronic equipment necessary to test the system, and we express our particular gratitude to the firm and all the gentlemen concerned. We thank the men and women of the Hollerith Department of the Max-Planck Society for doing the punch card work. This paper is dedicated to the Society's deceased director, Grad. Eng. Willi Heimerdinger. Our special thanks goes to the two gentlemen and institutes which made our work initially possible: Prof. Dr. Reinhold von Sengbusch and the Max-Planck Institute for Plant Cultivation [Max-Planck-Institut für Kulturpflanzenzüchtung], with which our documentation center is connected, as well as Dr. Martin Cremer and the Institute of Documentation, which has financed our work since 1 January 1962.

2. Definition of Technical Terms Used[1]

Terminology principle. We distinguish two principles for information retrieval methods. When applying the document principle, one data carrier (index card, punched card, magnetic tape section),

---

1. Terms, which are assumed to be generally known, as well as terms which are explained within the text, are not defined here.

containing all of the terms occurring in the respective document, p. 3 is used for each individual document. When applying the terminology principle, it is the other way around. In this instance, a data carrier is set up for each individual term, and the numbers of all documents which contain the respective terms are listed on the data carrier.

Current Contents. This is a literary service published in the United States, which gives the table of contents of certain groups of periodicals, even before their current issue is published, listing authors, article headings and addresses of the authors.

KWIC Index. This is a system developed by H. P. Luhn (USA). Its expanded name is Keyword-In-Context Indexing (or Index), and it alphabetically sorts (keywords) by machine public ion titles according to keywords contained therein and lists them in context with the titles.

Machine punch cards. Punch cards which are processed by machine. The cards are processed by machines individually (one after the other). They are not punched before their use.

Needle punch cards. These are punch cards which are processed mainly manually with needles or needle instruments. A number of cards can be processed simultaneously. They are already punched before use.

Notation. Any type of designation or symbolic representation, whereby other symbols take the place of the original words.

Topic tag. The shortest possible linguistic expression for the contents of a publication or the contents of a part thereof. The topic tag, therefore, refers basically to the contents of a publication. This distinguishes it from a key word, which is taken from the title of the publication in question.

Science Citation Index. A method developed in the United States, where bibliographic information about publications is the focal point of the documentation process. Proceeding from a specific

work and its author, a determination can be made, in particular, p. 4
as to which subsequent authors have cited this work in their publications.

Thesaurus. Any vocabulary wherein words are grouped according to related meanings.

3. The Basis for Our System

Why only titles?

The limitation of our system to publication titles constitutes the major objection raised by critics. However, this criticism can be met by various important arguments, which justify our system.

First of all, titles are generally better than is assumed by widespread prejudicial opinion. A thorough examination of biological articles was conducted in the United States, according to which from 50 to 70% of the titles render the essential content of the articles in topical form.

The rapid expansion and popularity of the American KWIC Indexing method and the "Current Contents" Service proves that titles alone are of considerable use to the subscribers. Both projects are based exclusively on the analysis of titles, and it is of interest to note that they have contributed to an improved rendering of titles by authors.

Let us now consider the necessary expenditures for purposes of comparison. A tenfold increase in personnel and an approximately twentyfold increase in personnel fund expenditure would be necessary for contents analysis by key phrases or abstracts, taking our extensive automation into consideration, as compared to processing titles alone. This increased fund expenditure is caused by the fact that contents analysis by key phrases or abstracts can neither be carried out automatically nor can it be done by assistants. Experts are needed for this type of work, and they cost more money.

The use of experts causes additional problems. One of these problems is the well-known personnel shortage, which affects documentation to a greater extent, because the scientists in question initially take up research or enter other professions. In addition, it is difficult to coordinate a larger number of reviewers in such a way that a working method is achieved which is as uniform as possible. p. 5

Finally, there is a technical-organizational factor, which causes limitation to titles appear in a particularly rational light. Each documentation method necessitates taking over specific information concerning the various publications by copying from the original literature, whether it concerns simple card indexes, needle punch cards, machine punch cards, or electronic installations. This is the only way to establish a documentation service. Moreover, all documentation stages, from pure title processing up to evaluation of data and values, must be based on a common minimum selection of such information. Author, title, periodical, volume, year of publication and pages, as well as in some circumstances additional information about number of literature references, figures and tables, etc., are classed among these important data concerning articles from periodicals, with which we are mainly concerned here. Since all this information, of which the title is a part, has to be copied anyway for each publication, a method, which works exclusively with this information and which is based on the process of simply copying from the original material, offers the most favorable solution. At the same time, one must provide for leaving the way open to the higher documentation stages and no changes in the system become necessary if additional desiderata are forthcoming at a later time.

Why punch tape typewriters?

The punch tape technique offers a method which permits us to store a text, which has to be written anyway, during the actual

writing process for any desired type of further processing. This is done with punch tape typewriters. They work like normal electric p. 6 typewriters and at the same time punch all written symbols in a paper tape in accordance with a special code. This punch tape represents the memory. All information punched on this tape can be reproduced by machine at any time, and can be transferred by means of a punch tape typewriter reader onto normal index cards or needle punch cards as many times as desired, or it can be transferred to machine punch cards or magnetic tapes by using special converting machines. Thus, any desired, subsequently processing documentation system can be accommodated. The punch tape typewriter makes it possible to perform the entire process of indexing information using assistants, who only have to copy all the informational data mentioned above. A good production output can be achieved by rational organization of the work. A group of four typists can, based on our experience, take down and correct an average of 100,000 bibliographic units a year with <u>one</u> punch tape typewriter operating several shifts a day, taking into consideration vacations and other normal absences.

<u>Why electronic computers?</u>

Electronic computers are the most technically advanced systems which have been used to date for documentation purposes. They alone enable us to completely automate all bibliographic work, to the extent that it is based exclusively on the evaluation of recorded titles and other information mentioned, and to make exhaustive use of the stored material in every respect possible.

The most difficult partial task within the framework of such total automation is the development of an automatic classification of publication titles. However, after this problem is solved, the p. 7 electronic machines can relieve man of having to compile title bibliographies and search through titles for specific subjects.

It is difficult to estimate how much outlay could be saved thereby, because to date incalculable double and multiple work is being done in this respect. Innumerable librarians, documentation specialists and scientists scrutinize the same literary material independently of one another, in order to compile bibliographies or find specific works. On the other hand, we record all original publications from all pertinent periodicals only once, and then let the machines do all other work automatically. It must be also taken into consideration that the print-out of the desired information follows automatically at a high speed and completely error-free, and can be repeated as many times as desired, so that savings result in this respect also.

Modern installations can also solve a number of problems, of which we will mention only a few at this point: compilation of annual publication lists (or lists extending over several years) of individual periodicals; compilation of alphabetical and systematic indexes; compilation of catalogs for libraries, institutes and individual scientists; compilation of systematic vocabularies (thesauri); general literary research with inherent possibilities for determining gaps in research, classifying periodicals and investigating the development of scientific documentation in general.

II. THE CLASSIFICATION SYSTEM

Automatic classification and, as a prerequisite therefor, a suitable classification system constitute the focal point of our entire system. Work on it has been going on since 1952.

1. General System Types p. 8

There are two major types of classification systems: hierarchical systems and basic term systems, which can be combined arbitrarily.

a) Hierarchical systems are strictly classified in a step sequence of categories. They start with a very general, and end with a very specific term. Their logically unobjectionable application

is given only if a certain number of subjects are classified according to a continuous and standard point of view (basis for classification). One example is the phylogenetic system of organisms. Here, the natural relationship is used as the basis for classification.

b) Basic term systems, which can be combined arbitrarily, forego hierarchical classification. They are based on the concept that category steps can be exchanged, particularly when classifying characteristics, so that each term can be arbitrarily used as a broader or narrower term. Let us mention the terms "blood" and "temperature" as an example. "Blood" can be the broader term for all its characteristics, as for instance "temperature," "color," and "viscosity." But "temperature" can just as well be used as a broader term fo. its occurrence in the form of "blood temperature," "body temperature," "water temperature," "soil temperature," etc. Basic term systems, consisting of terms which can be arbitrarily combined, have found repeated application in documentation recently. They are used to take the numerous relationships between terms into account to the maximum possible extent.

c) Both systems have their specific advantages and disadvantages. The disadvantage of the hierarchical system consists mainly in its inflexibility, and the fact that when using it for characteristics, the same terms are repeated at various points in the system. For example, the term "temperature" appears 44 times in the international Universal Decimal Classification System (UDC). The disadvantage of terms which can be arbitrarily combined lies mainly in the fact that broader terms are lacking, since a minimum p. 9 of hierarchy is indispensable for the formation of term groupings.

2. Own System

We have taken the following course in our system of natural sciences, wherein emphasis is on the field of biology:

a) subjects and characteristics are strictly separated. Chemical compounds and organisms are classified as subjects. All other terms are taken as characteristics;

b) pure hierarchical systems are used for the subjects. We developed a mixed system for the characteristics, which combines the advantages of both classification principles (hierarchy and arbitrary basic term) and avoids their disadvantages;

c) the basis for classification of chemical compounds is the type and number of atoms present in the compounds. Expressed another way, the compounds are categorized in accordance with their gross formula, which at the same time is used as the notation. It is structured as follows: ZO stands at the beginning as a standard symbol for all chemical compounds. These symbols were selected because Z does not occur at the beginning of any other notation and the numeral O separates this letter [Z] from the subsequent element symbols. Then follows the gross formula, where the indices appear as normal numerals. In addition, atoms occurring only once were given the numeral 1. For example, sodium sulfate (Na_2SO_4) receives the notation $ZONA_2S_1O_4$.

Such regulation is necessary in order to avoid errors during machine processing and print-out. The machine receives the command to consider the one or two letters which stand between numerals, as element symbols, while the numerals itself reflect the number of respective atoms. Since these notations can only be printed in capital letters, the machine would not be able to differentiate between the symbol for zinc (Sn) and the symbols for sulfur (S) and nitrogen (N), without the use of the numeral 1, for example;

d) the phylogenetic relationship is used as the basis for classification for the organisms. The natural or phylogenetic system of organisms, which has this classification basis, has two serious disadvantages as compared to the chemical compounds system. First of all, no fixed and generally acknowledged system

p. 10

exists, because the phylogenetic relationship cannot be determined directly and clearly, as is generally the case for the structure of chemical compounds. The natural relationship must rather be found indirectly through biological research. This results in numerous distinct opinions about the correct classification of organisms, and in addition, the system changes with the progress made in science. Second, there is no notation system for the organisms which is even approximately equal to chemical formula expressions and which reflects the relationships throughout all category steps.

We had to eliminate both disadvantages, because a fixed classification system, as well as a suitable rendering of the notations was necessary for documentation and particularly for our method.

We therefore decided on a classification system which followed the best known and most thoroughly systematic works and digests, but which took into account only the most common category steps. We developed a notation system composed of numbers and letters for this purpose, the notations of which reflect the entire relationship of the individual organism groups;

o) the characteristics system contains four groups of systems having 10 individual systems each. There are, therefore, a total of 40 individual systems. The notation system in question consists of four-digit notations. The first digit indicates the system group and the second digit indicates the individual system. The third and fourth digit are used for individual basic terms.

Thus, we have a classification with four gradations. This is the minimum hierarchy, which has been tested by us and found satisfactory for a usable system.

Category	Notation Place	Symbol
Division	first place	letter
Class	second place	number
Order	third place	letter
Family	fourth and fifth place	numbers
Genus	sixth and seventh place	numbers
Species	eighth and ninth place	numbers

Example: House dog (Canis familiaris)

Notation R9H0101O2

Chordata: R

Mammalia: R9

Carnivora: R9H

Canidae: R9H01

Canis: R9H0101

House dog: R9H0101C2

Each basic term appears only once in the system. All basic terms can be freely combined with each other and with the terms of the subject systems. Pertinent broader terms appear automatically in connection with each combination of basic terms because of the four-gradation hierarchy. This results in an optimum supply of information.

The following is a survey over the system groups and individual systems of the overall system:

0 Major categories

00 Characteristics carrier*

01 Overall scientific classification

02 Methods

06 Organizational steps

07 Cellular components

08 Classification of the entire organism

---

* Translator's Note: The word appears in the text as "Semaphoront," and is further on in the text defined as "Merkmalstraeger," a characteristics carrier.

03 Generation

04 General broader terms

05 Individual scientific fields

09 Various functional terms

1 Basic scientific categories

10 General basic terms

11 Magnitudes

12 Rhythms

13 Medium and condition

14 Mechanical factors

15 Energy factors

16 Chemical factors

17 Major groups in cyclic system

18 Sub-groups in cyclic system

19 Organic-chemical factors

2 Individual categories of idiobiology

20 Nervous system

21 Digestive system

22 Urogenital system

23 Attack and protective system

24 Circulatory system

25 Supporting system

26 Motor system

27 Sensory system

28 Respiratory system

29 Various systems

3 Individual categories of coenobiology

p. 12

30 Disaccharids: social relationships

31 Disaccharids: material

32 Coenocytes: land (lithosphere)

33 Coenocytes: air (atmosphere)

34 Coenocytes: water (hydrosphere)

35 Geography: planet as a whole

36 Geography: Europe

37 Geography: America, Arctic, Antarctic

38 Geography: Asia, Africa

39 Substrate

The overall system is set up with particular emphasis on biology in its widest sense, and on the earth sciences. However, it also contains basic physics and chemistry terms, and can be arbitrarily expanded and extended to other fields in accordance with its general principles.

All basic terms, which would be outside the principle of the four-gradation hierarchy if they appeared in one of the other individual systems, are contained in system group 0. These terms are therefore compiled in a special system. In other words, they are major broader terms or basic terms which have a very general meaning. Let us take as an example the individual systems "characteristics carrier," "methods," and "cellular components."

The development stages, sex, and anomalies, as well as the systematic units below the species, are compiled under the broader term "characteristics carrier" [Semaphoront] (Hennig, 1950). The corresponding basic terms of this system can be taken as further classification of the organisms subject system, as well as in the sense of broader terms for all biological individual systems of the other system groups. Consequently, the characteristics carrier system occupies an intermediate position between the subject system and the characteristics system.

The individual system, "methods," shows in a particularly distinct manner that here a general, own point of view is present, which can be significantly combined with any other basic term of the other individual systems.

"Cellular components," finally, form examples of basic biological terms, which would be repeated in all individual systems of group 2 because of their common appearance, if a special individual system had not been created for this purpose within the framework of the main categories. p. 13

System group 1 consists chiefly of the basic terms occurring in physics and chemistry.

The individual organic systems are found in system group 2. The advantages of the principle of combining basic terms can be especially well demonstrated with the pertinent terms of this system group. Let us take as an example the term "sense of light"

(notation 2730). This term, as all the other basic terms, is to be understood in its most general meaning, standing alone among the terms, as it were. It obtains concrete meaning only by a combination with basic terms, particularly of system group 0; let us illustrate this with the following examples:

2730 + 0601 (excreta): lacrimal fluid

2730 + 0610 (cell): visual cell

2730 + 0620 (tissue): eye tissue

2730 + 0630 (organ): eye

2730 + 0640 (entire organism): eyesight

or

2730 + 0630 + 0521 (morphology): eye structure

2730 + 0630 + 0532 (developmental physiology): eye growth

2730 + 0630 + 0533 (genetics): eye heredity

2730 + 0630 + 0542 (evolutionary physiology): eye mutation

The combination of few, well-founded and well-defined basic terms from few individual systems results in a multitude of concrete terms, as is illustrated in the above examples, which in order to save space are incomplete here. Taking into consideration in addition the numerous synonyms, it is then understandable that we were able to clearly define the words and terms or the characteristics system (approximately 50,000), which have appeared during the processing of approximately 40,000 files, by combining only about 2,000 basic terms.

System group 3 consists of individual categories of the field of coenobiology. While idiobiology (system group 2) concerns itself with individual organisms, all those relationships are compiled under coenobiology, where individual organisms occur as subgroup of a higher category. These higher categories consist either of individual organisms (for example, parasitic or sexual relationships) or of organismic and non-organismic subgroups. This latter

p. 14

is valid for coenocytes, which are defined here as a unit of living space (non-organismic) and living community (organismic). The geographic basic terms are used accordingly. Therefore, all terms from the earth sciences (geography, geology) can be expressed by our system.

III. THE SYSTEM

The system can be divided into three sections, which follow one another during the sequence of operations: indexing of bibliographic units including transfer to magnetic tape; automatic classification including all operations necessary for it; the processing, which varies according to the purpose for which the information is needed, and output of the material.

1. Indexing and Transfer

a) Procedure for periodicals: recording the title and other bibliographic data of all original publications of as many periodicals as possible with punch tape typewriters. A fixed model is needed for this purpose as for all machine work. This model has a category number for every information category, and a line number within every category for each indexed line, so that a two-digit number stands in front of each printed line. (See the example in Chapter III/2/e). Individual categories are divided as follows:

Category 0: publication number

It contains first of all a five-digit number for the periodical with a capital letter added as prefix. The letter is used for p. 15
the classification of the periodical according to its technical field. In other respects, all periodicals are numbered consecutively without special sequence. Next follow three digits which are used for stating the volume of the periodical. A last three-digit or four-digit number designates consecutively each individual article within volume and periodical. - A completely clear designation of each work is obtained with these publication numbers as a result.

Category 1: intended for corporative author (institution).

Category 2: author.

Category 3: original source.

In this case, the volume number stands before a parenthesis, in parenthesis is the year of publication, and after the parenthesis the beginning and final page of the article. The name of the periodical does not have to be recorded in this category, when indexing the bibliography, with which are are concerned at this point. The machine stores all names of periodicals separately under their periodical numbers (see category 0) and prints them automatically in front of the volume number during print-out of the bibliographic data.

Category 4: notation of supplements and other.

The following belong to this category: number of bibliographic quotations, figures and tables, as well as language of the article, and existence and language of abstract in another language. In addition, the machine computes from beginning and ending page (see category 3) the size of the article for print-out and prints this information as the first number in front of the other information.

Category 5: intended for secondary sources (references in a publication of references).

Category 6: title (with subtitle) and translation of title, if needed.

Category 7: topic tags or abstracts.

Category 8: intended for the Science Citation Index system.

Category 9: available for supplementary remarks.

All this information is typed onto a DIN* A6 index card for each publication. Index card-continuous forms are used in order to avoid as much as possible having to insert the cards

p. 16

---

* DIN - German industrial norm.

into the typewriter and taking them out again. A tape is punched simultaneously as the cards being typed. The index cards are used for control purposes later on.

b) Transfer of the information from the punch tape to magnetic tapes using the IBM 1401 computer and the IBM 1011 sensor connected to it. The print-out of lists of corrections to the indexed material using the IBM 1403 printer.

c) Correction of errors made during input and transfer of data.

2. Automatic Classification

a) Retrieval and counting of words appearing in the titles (data category 6) of the publications with the IBM 1401. Consolidation and output of words in the form of punched vocabulary cards. A vocabulary list stating frequencies and where word was found (word frequency list) is put out at the same time.

b) Classification of the words listed on the vocabulary cards and handwritten entires of appropriate notations on these vocabulary cards.

The classification work at this point in the work sequence has to be carried out only once for each word by the human processor. The more words and corresponding notations the computer has stored in its internal vocabulary, the less is the number of added words. Some day a stage should be reached, where only words, which are newly introduced into science will be added.

The words are initially divided into seven different vocabulary categories: p. 17

expletives (1)	name of species (2)
chemical compounds (3)	name of organisms (4)
terms (5)	nomenclature/term (6)

first names (7)

All those words, which do not have any informational value, are categorized as expletives. (Examples: the, and, or, investigations, results).

The name of species, belonging to biological nomenclature, are compiled in a separate vocabulary category. This is necessary because for one the names of species are isolated words, just as all other words, and for the other clearly designate the type of organism in question only in combination with the pertinent generic name. The special solution of the problem resulting therefrom will be described in the section concerning output (examples: palustris, Hustedti).

Vocabulary category 3 contains all chemical compounds which can be given for a gross formula (example: sodium hydroxide, barbituric acid).

All scientific designations and all trivial names for organisms of all categories from division to genus, as well as trivial names of species, are included in the category of organism names (example: plants, birds, Marchantiaceae, Drosophila, deadly nightshade).

All words in the characteristics system are grouped under terms (examples: albino, immunity, gravitational effect, stomach nerves, Central America).

In vocabulary category 6, we find all designations made up of organism names and characteristic terms, and all other words in which organisms and characteristics notations occur together (example: bacteria cell walls, amphibian ova, blossom colors).

A special vocabulary category was created for the first names of persons, which will not be considered further in the following description. It became evident that the articles in question concerned, for the most part, obituary or other biographical articles whenever these words appeared in the titles of publications (example of title: Karl Lueders, 70 Years Old).

Expletives (WB [vocabulary category] 1) and names of species (WB 2) receive the vocabulary category only, and do not have any notations. Chemical compounds (WB 3), names of organisms (WB 4)

and combination words (WB 6) are classified according to the subject system in question and tagged with the corresponding notation.

Terms (WB 5 and combination words WB 6) are classified in accordance with the principle of definition. The meaning of p. 18 the word is expressed by combining the notations of basic terms present in the characteristics system, with a notation from the subject system added, if needed. For example, the word "pollen" can be defined as "male (notation 0092) germ cells (notation 0000) of gymnosperms (notation G) and angiosperms (notation H)," and represented by the appropriate notations.

All synonymous words in the same language and corresponding words of other languages receive the same notations.

Furthermore, each word receives a language symbol, so that later on all words of the same language can be collated by machine. A polyglot vocabulary can be obtained in this manner.

c) Punching the vocabulary category, the language symbol and the notations on the vocabulary cards. The vocabulary cards are subsequently fed back into the electronic computer. A thesaurus is created on the magnetic tape in this manner, which can be continuously supplemented with new words and words which are not yet recorded on the tape.

d) Internal machine verification and sorting processes. Print-outs of alphabetical and systematical vocabulary lists (thesauri). Additional proofing and correcting processes.

e) Basically, two "clean" magnetic tapes are available after all verification and correcting work is finished. One tape contains the input of bibliographic units with all pertinent data, and one tape the thesaurus (words and notations). The computer carries out the automatic classification with these two tapes and by using a few additional internal magnetic tapes. In accordance with the terminology principle, it simultaneously constructs two registers which serve to accelerate interrogation.

By way of example, we will show three bibliographic units, as well as a retrieval sample from the alphabetically structured thesaurus, in order to illustrate the work sequence in a simplified, easily understandable manner. p. 19

Bibliographic units

The input model is explained in detail in Chapter III/1/a. Numbers in parentheses after each word are word numbers and only serve explanatory purposes.

00 A00011-046-021
20 Mutschler, Albert
30 46(1954) p. 364-418
40 O-3-1-Germ.-Engl., Russ.
60 Mass occurrence (1) of(2) larvae(3) of(4) Eristalis(5)
61 tenax(6) in(7) waste water(8).

00 B00104-085-006
20 Steiniger, Otto Hellberg, Karl-Heinz
30 85(1936) p. 12-29
40 15-O-2-Germ.-O
60 The(1) change(2) in(3) leaf size(4) and(5) color of blossoms(6)
61 of frequently(7) cultivated(8) flowering plants(9) after (10)
62 fertilizing(11) with(12) waste water(13) containing phenol(14).

00 A00117-008-019
20 Koester-Uhlig, Heinrich, von
30 8(1959) p. 14-33
40 31-O-3-Germ.-O
60 The(1) size(2) of(3) chromosomes(4) of(5) Diptera(6).

Thesaurus extract

Word	Vocabulary category	Language symbol	Notations
Waste water	5	B	1311
Change	5	B	0420 1019
Cultivated	5	B	0220
Leaf size	5	B	0020 2950
Color of blossom	6	B	H 1530 2270
Flowering plants	4	B	G H
Chromosomes	5	A	0790
The	1	B	-
The	1	B	-
Diptera	4	A	09L
Fertilization	5	B	0226
Eristalis	4	A	09L2109
Size	5	B	1120
Frequent	1	B	-
In	1	B	-
Larvae	5	B	0030
Mass occurrence	5	B	1115
With	1	B	-
After	1	B	-
Containing phenol	3	X	Z0C6H601
tenax	2	A	-
And	1	B	-
Of	1	B	-

The computer provides all words in data category 6 (titles) p. 20 of each bibliographic unit with a consecutively running number during the first work sequence (in our example, the number is after each word in parentheses). The machine then compares these words with the thesaurus and assigns to each thesaurus word the

publication numbers (data category 0) of all bibliographic units, where the word occurs in the title. The consecutively running word numbers are added to these publication numbers at the same time. Finally, two magnetic tape registers - a word register and a notation register - are generated on the basis of this intermediate internal computer stage.

The word register no longer contains the expletives (vocabulary category 1), since they were eliminated by the computer. The word register is arranged alphabetically and consists of all declarative words with corresponding publication and word numbers.

Vocabulary category 2 (names of species) is eliminated in the notation register, because notations are not needed for this category. The notation register is sorted according to notations and thereby takes on the characteristics of a systematic register as compared to the alphabetical word register. Corresponding publications and word numbers are again listed with the individual notations. The automatic classification process is finished with the generation of the notation register.

The advantages of this method are clearly evident. The automatic classification is done more than a thousand times faster than it could be accomplished by man, and in addition, operates in a completely uniform and error-free manner. Moreover, all publication numbers on one subject are already in one group (terminology principle) so that the machine does not have to search through all the stored bibliographic material for the inquiry operations, which will be described later.

3. Processing and Print-Out

Computer processing of bibliographic material for the various kinds of evaluation and output, some of which we have listed at the beginning of this article. Subsequent machine retrieval of digests, lists or any of the index cards in accordance with the purpose for which they will be used. p. 21

We will describe the evaluation and print-out process for documentation inquiries as a final example. For this purpose, it is best to look at our entire system from the standpoint of the user. After all, contact with the user is of primary significance, since it is the task of documentation to be of service. Subsequently, we will list some principles which are of importance to users of documentation and then cite some inquiry examples.

a) Basically, the user should be completely free in formulating his inquiries and should not be obligated to keep to any pattern. It is the task of the documentation centers to rephrase the user's inquiries for the computer. Personal cooperation between user and documentation center is recommended for difficult or vague inquiries.

b) The computer can answer all questions relating to data stored for each bibliographic unit and any of their combinations, which has been described in detail in Chapter III/1/a. An extreme example, which will hardly come up in practice, would be the question, "All publications by G. Kuntze from the Genetics periodical covering the years 1950 to 1960 concerning the heredity of eye color in man, with less than 20 pages, more than 10 bibliographic data, 5 figures and 5 tables, in German, with an abstract in English." - It is more efficient for a rational utilization of the computer to process several hundred inquiries simultaneously.

c) According to the results of an inquiry test, which was carried out by us and answered by more than 350 future users, our system will have to fulfill two major tasks within the framework of documentation inquiries: furnishing by subscription, about once a month, all publications of the recorded periodicals pertaining to a specific subject; and replying to single inquiries on the basis of all stored bibliographic material. p. 22

d) Inquiries by subscribers are usually concerned with the particular direction of work the inquiring institute or scientist is taking. Examples for such inquiries, taken arbitrarily from the inquiry test, are: "Salmonella biology," or "Animal cell differentiation," c. "Everything concerning the Centaurium genus," or "Human ethnology," or Nucleoles in plants and animals," or "Electronic microscopy in biology and medicine." All these cases indicate that one wants to stay informed about one's own field. Our system is arranged in such a way that either more condensed or more extensive material can be supplied if desired in answer to subscribers' inquiries, since library experiences have taught us that those publications which are at the periphery of the actual research subject, can also be important.

e) The possibilities for single inquiries became crystallized, in particular, during the development of our system. Since entire periodicals have been processed, even back volumes, a determination can be made as to whether or not earlier publications dealing with a specific problem are available by directing a single inquiry to the total volume of our bibliographic store. Thus, the entire store of knowledge which has been indexed by us, can be made available to the user, so far as it is expressed in the titles of the publications. Such questions could be, for example, "Where is a method describing the measurement of slit apertures?", or "Has anybody worked on saliva secretion of Phylloxera vastatrix as dependent on changes in daily atmospheric pressure?", or "Are any investigations concerning type-specific differences in erythrocyte content of oestradiol-dehydrogenases in mammals available?", or "Has the presence of the cabbage moth as a pest on poppy ever been determined?", or "Has a comparison been made between the albumin percentage of the tobacco mosaic virus and that of a temperature mutant?" - These examples stem from our biblio- p. 23
graphic material. Even if they exist, no standard bibliography

or register could supply the answers. The computer, however, can accomplish it. As soon as one pertinent work has been found by the computer, additional information results (most of the time) from its bibliographic data.

f) In conclusion, here are some examples of computer replies to inquiries. We refer back to Chapter III/2/e and the examples cited there for this purpose.

Let's assume that a user poses the question: "All work concerning waste water problems." The pertinent notation 1311 is then put into the electronic system by punch card. The computer looks in its notation register for the notation in question and there finds all publication numbers of articles on the subject. With the aid of these publication numbers, the machine retrieves all appropriate titles and other bibliographic data from the bibliographic magnetic tape and prints the data in any sequence and classification desired on lists or index cards.

Notations 0420, 1019 and H, 1530, 2270, are put into the computer for an inquiry "color change in the blossom." At the same time the computer is told by special symbols via the appropriate punch card, that both notation groups are to be taken as a logical product and that these notation groups represent a notation chain. The logical product signifies that the notations in question must occur in combination. A notation chain is present whenever several notations belong to one word or term, which are to remain together in the specific case and cannot combine with other notations of other words of the same title and thus form a wrong statement. Then the computer uses the word numbers. In our example, it retrieves the notation for"color of blossom" (H, 1530, 2270) and compares all publication numbers including their word numbers listed under these notations. The computer finds that only number B00104-085-006-(6) is identical in all three notations. This number is then compared with the publication numbers listed among notations for the word "change" (0420, 1019), whereby it is found that the

same number appears there also. We therefore have a paper concerning "change in the color of the blossom" with publication number B00104-085-006, which is then printed out accordingly.

The computer searches for notation 09L for an inquiry concerning all articles about Diptera. It not only retrieves all those articles stored for the notation 09L itself, but all other articles of all notations starting with 09L, for example the notation for Eristalis (09L2109). This example illustrates the fact that all narrower terms (in our example the families and general of Diptera) are automatically included in our method of giving notations. p. 24 It would necessitate searching for several hundred or even thousands of different names in a regular alphabetical register, which practically no one could accomplish.

An inquiry concerning a specific type of organism, as for example Eristalis tenax, requires a special procedure, because the name of the species (tenax) has the desired informational value only in connection with the generic name (Eristalis) (cf. the description of vocabulary categories). In all such cases, the computer searches for the genus in the notation register, and for the species name in the word register. Only trivial names, as for example, "cockchafer" [a large European beetle], consisting of only one word, have a notation going as far as the species, and therefore can also be found in the notation register.

The punch cards, onto which our various inquiries are punched, are kept, so that an inquiry index is being formed over a period of time. Incoming new inquiries are first checked to see if they are already in the index, so that uncessary work can be precluded.

IV. CONCLUDING REMARKS

1. Limitations of the System

It would be unscientific to remain silent about what our system cannot accomplish and what its limitations are. The

reasons for the limitation to titles of publications have already been cited in detail above. This ties in with the fact that it is not possible to answer inquiries pertaining to the results of concrete research, such as: "How many chromosomes does the sunflower have?", or "How long are the intestines of a guinea pig?", or "Which plants have yellow flower petals and bloom during the first four months of the year?" In order to be able to do so, it would be necessary to develop the highest documentation stage (the documentation of findings), which calls for totally different prerequisites. Furthermore, it is in the nature of things that automatic classification also presents a series of problems, the detailed description of which is not possible here. These are by and large linguistic problems, similar to those which confront machine translation. Problems of homonymy and questions of syntax, for example, belong in this category. We have taken these things into consideration during the development of our system to the extent that it was possible to do so in advance. In addition, practical work and continuous information activity p. 25 will give us the chance to eliminate shortcomings which still exist in cooperation with the users. Work on the classification system will continue at the same time, because the system allows us to reclassify the material by machine, if necessary. Despite the remaining limitations, which are conditioned by the structure of the language, we are of the opinion that we can continue to help users considerably with our system. While so doing, we must remember not to put the cart before the horse, and that 90 or even 100% solutions are not possible anyway in the field of documentation.

2. Outlook

The method has passed its trial test in principle, and thanks to the efforts of the Institute for Documentation at the Max-Planck Society in Goettingen an electronic installation is available

exclusively for documentation purposes, and we hope to be able to expand the application of the system to all scientific fields and create a machine-processed "National Bibliography" of periodical articles published in the German language.

We would then have a broad foundation for the automation of publications, libraries and documentation, which could handle all possible applications indicated at this time. At the same time, it would be desirable to have appropriate classification research carried out with this material. This is a field which is steadily gaining in importance on the international level and which should also gain acceptance at universities.

The "National Bibliography on Magnetic Tape" should not represent any competition to existing documentation centers and for this reason should not supply direct information to users. This would require much too large an organization; the organization already to be found at proven operational documentation centers is large enough. The best solution, therefore, seems to be that p. 26 each documentation center receive the material which is of importance for that particular center, in reproduced form, from the National Bibliography Center. The individual centers can then continue to work with the material in any way they desire.

In the future it may be possible to offer the national bibliography as a whole - as a magnetic tape duplicate - to foreign countries on an exchange basis, in order to stimulate corresponding documentation projects in other countries. It could thus come to pass that there could even be international regulation of the first documentation stage of indexing and evaluating titles in the not too distant future. Further work could then develop on the basis of this regulation.

Literature

Only a few general works, which are connected with the structure of documentation in general, or our documentation center in

particular, are listed here for reasons of space. Persons looking for specific technical literature are referred to the bibliography [4] and the periodicals on this subject (in Germany, particularly the Documentation News).

1. Cremer, M., "Task and Function of the Institute of Documentation." Max-Planck Society Reports, 1963, No. 1/2, pp. 77-86.

2. Pietsch, E., "Structure of Information Systems." Documentation News, Vol. 15, No. 1, pp. 28-41 (1964).

3. Scheele, M. Punch Card Methods in Research and Documentation, With Particular Consideration Given to Biology. 2nd ed., Stuttgart: Schweizerbart, 1959.

4. Scheele, M., Literature Concerning Punch Card Methods. A bibliography concerning punch cards decoded with visual punch cards, and problems connected with its application. Schlitz/Hessen: H. Guntrum H.K.G., 1959.

5. Schneider, K., "Five Years of KWIC Indexing According to the H. P. Luhn System." Documentation News, Vol. 14, No. 4, pp. 200-205 (1963).

6. Sengbusch, R. v., "Problems of Selection and its Relationship with Our Cultural Life." Max-Planck Society Yearbook, 1959, pp. 145-185.

Max-Planck Institute for Plant Cultivation, Hamburg

Submitted on 19 July 1964.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author) Translation Consultants, Ltd. 944 South Wakefield St., Room 302 Arlington, Va. 22204	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE "An Automatic Classification System for Publications, Libraries and Documentation." (EIN VERFAHREN ZUR AUTOMATISCHEN KLASSIFIZIERUNG FUR VEROFFENTLICHUNGSWESEN, BIBLIOTHEKSWESEN UND DOKUMENTATION.)

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Translation

5 AUTHOR(S) (First name, middle initial, last name)
Von Martin Scheele, Schlitz/Hessen

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
1968	30	6

8a. CONTRACT OR GRANT NO. N62306-69-M-0201 b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) NOO CONTRACT TRANS 12

10. DISTRIBUTION STATEMENT
Each transmittal of this document outside the agencies of the U. S. Government must have prior approval of the Naval Oceanographic Office.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
Reprint from the Periodical NATURAL SCIENCES	U. S. Naval Oceanographic Office Washington, D. C. 20390

13. ABSTRACT

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Security Classification

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Electronic Computers Documentation Information Retrieval Data Processing Information Systems						

DD FORM 1 NOV 65 1473 (BACK)
(PAGE 2)

Security Classification